FILE FORMING SYSTEM FOR CUTTING OF NEWSPAPER AND MAGAZINE

Publication number: JP57199066

Publication date: 1982-12-06

Inventor: SAKAI TOSHIYUKI; INAGAKI KOUSAKU

Applicant: SAKAI TOSHIYUKI

Classification:

- international: B42D7/00; G06F17/21; G06F17/30; G06T1/00;

G06T11/60; B42D7/00; G06F17/21; G06F17/30;

G06T1/00; G06T11/60; (IPC1-7): B42D15/00; G06F3/14;

G06F15/20; G06F15/40

- European:

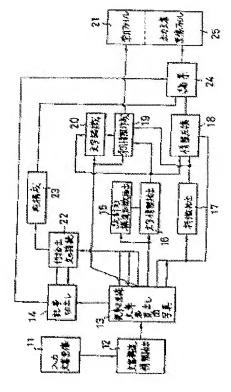
G06T11/60

Application number: JP19810084589 19810602 Priority number(s): JP19810084589 19810602

Report a data error here

Abstract of JP57199066

PURPOSE: To perform display or filing through the rconstitution of the interior by cutting and picking up the contents of a desired article from newspapers and magazines as an inverse processing as the cast-off jobs of newspaper edition. CONSTITUTION:An input document picture 11 being an objective to be processed is introduced to a document construction information pickup section 12, and after the architectual information is checked, the information is classified into article section lines, sentences, tables, index, graphs and pictures at an information sectioning section 13 and one set of article is cut out at an article cut-out section 14. The index information is discriminated for the shape by using the statistical architectual knowledge at a construction knowledge pickup section 15. The index character information is picked up at a character information pickup section 16 with the information from the pickup section 15, special features such as graphs and pictures are picked up at a character pickup section 17, and data such as contraction of size is compressed at an information compression section 18. The characteristic information such as the index, graphs and pictures is inputted to an index forming section 19 to form the index data and to obtain an index file 21. The main sentences of an article obtained at the section 14 is discriminated for the pickup of lines and connection of sentences at a structure recognizing section 22, and reconstructed 23 and edited 24, and registrated in an output



document	picture	file	25
----------	---------	------	----

Data supplied from the **esp@cenet** database - Worldwide

(19) 日本国特許庁 (JP)

⑩特許出願公開

⑩ 公開特許公報 (A)

昭57-199066

௵Int. Cl.³	識別記号	庁内整理番号	砂公開 昭和57年(1982)12月6日
G 06 F 15/20	1 0 3	7157—5B	
B 42 D 15/00		7008-2C	発明の数 1
G 06 F 3/14		2116—5B	審査請求有
15/40		6913—5B	
			(全 16 頁)

砂新聞・雑誌等の切抜きファイル作成システム

②特 願 昭56-84589

②出 願 昭56(1981)6月2日

⑫発 明 者 坂井利之

高槻市奈佐原元町158番地

四発 明 者 稲垣耕作

京都市左京区吉田中大路町32番

地

⑪出 願 人 坂井利之

高槻市奈佐原元町158番地

⑩代 理 人 弁理士 鈴江武彦 外

外2名

明 細 苞

1. 発明の名称

新聞・雑誌等の切抜きファイル作成システム

2. 特許請求の範囲

文書画像情報として入力される新聞や雑誌等の紙面に記載された記事群の中から一つの境界を区面線または段間および行間スーた文章構造のつながりの関係から所望とする記事本文とその付随場を抽出し、この抽出した記事本文とその付情報とを紙面内での割付領域を表示したり、再構成処理して表示またはファイルで成システム。

3. 発明の詳細な説明

本発明は新聞や雑誌等から所望とする記事内容を効果的に切抜き抽出し、これをそのまま再編集して表示またはファイルすることのできる実用性の高い新聞・雑誌等の切抜きファイル作

成システムに関する。

新聞や雑誌等には数多くの記事が予め定められた枠組み構成規則に従ってレイアウトして戦せられている。そして、最近では、これらの中から所望とする記事を人間が切抜きファイルし、再構成して計算機システムにファイリングすることが試みられている。

この場合、新聞・雑誌等の紙面サイズが決められているものの、各種記事は上記紙面サイズを有効に利用すべく、或る領域に割付けがなされて組込まれているので、単に紙面上の単純な位置を特定するだけでは所望とする記事のみを切扱き抽出することができない不具合がある。また表のように枠の見出してある属性名を指定することによって記事内容を切抜き抽出するととできない。

つまり、1枚の紙面上で表示あるいはファイル化の対象となる記事は多数存在し、その中で種々の形状に割付けがなされた或る一つの記事だけが一つの単位記事となるので、例えば1紙

-- 1 ---

面をページ等とする固定的な単位のみで取扱りをといる。またコード化された文字情報を扱うを扱うをひて、文書画像として入力理解ので、必然的に文書内容の理が伴い、その結果パターン理解的な情報の一つの記事は記事本文のみによって開発を構成されるのではなく、含まれるから、ことが要求される。

本発明はこのような事情に鑑みてなされたもので、その目的とするところは、新聞や雑誌等の中から所望とする記事内容を新聞製作の割付作業の逆処理として効果的に切抜き抽出して、割付領域を抽出表示したり、その内部を再構成して表示あるいはファイルすることのできる実用性の高い新聞・雑誌等の切抜きファイル作成システムを提供することにある。

本発明は文部画像として入力される新聞や雑

— 3 —

て 構成され、従って 1 枚の紙面中には複数の記 事内容が混在したものとなっている。また記事 本文3の境界は凝微に付された区画線4や段間 あるいは行間スペースによって示されるように なっている。しかし一つの配事においては11 2,3は隣接関係を保ってまとまるようになっ ている。そして、記事本文は、上記区面された 境界内において、例えば凝むき日本語文章の場 合、右から左へと行を移動し、最終行からはそ の下段の棚に移って同様に右から左へと行を移 動して記載されるようになっており、1つの記 事本文のは、一般に或る領域内にまとめられる よりになっている。しかして表示あるいはファ イル化の対象は、とのような紙面中の或る記事 内容となるから、そのレイアウト形状は一定化 せず、他の記事内容との兼合いにおいて種々の 形状となる。従って、表示あるいはファイル化 の対象を紙面の中での単純な領域として固定化 した単位として扱うことは到底望めない。しか も画像として文書を処理するので、従来の文字 以下、図面を適宜参照して本発明の詳細を説明する。尚、ととでは新闻を例にして説明するが、一般的な雑誌や都輸等も対象となるととは云うまでもない。

さて、新聞は第1図に模式的に示すように、 その紙面に見出し1、写真2、記事本文3を所 定の割付け規則に従って多段に亘って割付けし

-- 4 --

コード化された文書を扱うものとは異なり、あ ちゆる形態の文書情報に対処しなければならな い。この為、文書内容の理解が必要となってく るので、ことにパターン理解的な情報処理手法 の導入が必要となってくる。

そとで本システムでは次のようにして画像入力される新聞・雑誌等の文書情報から所選とする記事内容の切抜きファイルを行っている。

 その逆操作をパターン理解的に行うととによって、第2図(d)に示すように見出し画像中から見出し文字を抜き出すことが可能となる。従ってとのようにして見出し1を抜き出せば、 これをファイル化において情報圧縮することが可能であり、またその文字をパターンのままでマッチング処理の対象として索引情報に用いたり、更には文字認識して文字コード化して索引情報とすることが可能となる。

尚、この見出しからの文字情報の抽出は、例えば画像縮退(Shrin king)処理や画像融合(fusion)処理を用いたり、画像の白黒反転処理を適宜用いることによって実施できる。特に文字抽出された見出しの文字パターンのサイズを縮小すれば、容易に情報圧縮ができ、その効果は非常に大きい。

ところで、不定領域形状の記事本文3の切抜 きも次のようにしてペターン理解的に行われる。 この記事本文3の抽出には、その記事を作成す るに際して用いられた領域地割りの為の区画線

- 7 -

在すると認識することができる。この際、次段の文章の縦の区画線で区切られた文章の中で、 右端の行の1マス目が空白でないものを、その 上段の記事文章のつながりであると認識すれば よい。

尚、このようにして見出しや記事本文を理解 して切抜く場合には、その前処理として見出し 部・記事本文部・図・表・写真等の画案配列の 異質な領域を個別に認識し、これを切り分けて や行間および段間スペース、また紙面の組み方に関する先験知識、 先に抽出された見出しの情報等が利用される。特に地割りの為の区面設は記事の境界を定める上で重要を役割りを果している。従って、 との区面線の線外を抽出してとれを理解すれば、 記事本文の境界を示す重要な要素として採用することができる。 との線分の もとにして 閾値処理すること等によって実現される。

さて、記事本文が複数の段に亘って組まれてといる場合、文書画像をより詳細に認識する際の先院細に認識する知識をよる。との場合には、記事の先験右には、記事の光験右には、上からではいい。またたけは、立ちたとは、立ちには、対策による文章の継続性の判定は、或なによっている文章の左端行の最後尾が句点相当の大きでないとき、必ず次の段に記事の続きが存

- 8 -

おく必要がある。との構造情報の抽出には、例 えば

村尾充洋,坂井利之

" 文 む 画 像 の 梅 造 情 報 抽 出 "

情報処理学会第21回全国大会

7 H - 1 (1980年5月)

に示される処理を用いるようにすればよい。つまり、 画案配列のマルコフモデルにおけるランレングスによる統計的知識を利用して写真, 図等を解折して抽出し、 しかるの ち前記した個々の処理を実行すれば、見出し、 記事本文の効果的な切抜き抽出が可能となる。 との処理は後述するとストグラム情報を用いた処理の一部で代替できる。

尚、囲み記事の場合には、見出しが記事本文を分断する形で入ってくるので、記事切出しの手順は少し異なってくる。しかしながら、囲み記事は必ず矩形領域の形状をとり、かつ見出しはその中央付近にあるので、そのような枠組み構成に関する知識を用いることにより、切出し

- 9 -

が可能であり、 ことで述べた手法から拡張すべき点は少なく、 容易である。 その他の特殊な形式の記事も、 適宜それらの枠組み 構成に関する知識を導入することにより、 同様に切り出せる。

第3図は、上述した画像処理を行って記事内容の切抜きファイルを作成するシステムの一実施例構成図である。

は画像に関する数値的な扱であっても良く、更には文字認識された文字コード等であっても良い。

-- 11 ---

一方、前記記事切出し部14に得られた1ま とまりの記事本文は、構造認識部22にて行の 抽出、文の接続関係等が判定される。との認識 部22は記事切出し部14と協働して作用する もので、とれによって、不定形に組まれた或る 1つの記事本文が抽出されるととになる。そし て、とのようにして抽出された記事本文は、再 構成部23にて、出力画像として適した取扱い の容易な形式に再構成される。とのようにして 再構成された記事本文に対して、編集部24で は、との文章に関連する前記見出しや図,表, 写真等を付加して編集し、とれを出力文書画像 ファイル25に登録する如く構成される。勿論、 原紙面はそのまま画像ファイルとして登録し、 それに加えて原紙面でのとの記事に対する紙面 割付けの領域情報としても登録できる。また計 箕磯システムでは通常田力装置もファイルとみ

に関する構造的情報を得て、これに適したアルゴリズムを用いて、見出しの文字情報を抽出している。また文書画像中の図、写真等については、特数抽出部11にてその概略形状情報等の特徴が抽出されている。そして、このようにして特徴抽出された図、写真の情報、かよび前記見出しの文字情報は、情報圧縮部18に導かれ、サイズの縮小や文字コード化等の個々の圧縮でルゴリズムに従ってデータ圧縮される。

また見出しや,図、写真,表等の特徴情報は索引アータとして索引作成部 1 9 に与えられる。また見出し及び表中に示される文字情報は、文字認識部 2 0 にて既性名を示す文字情報として、同様に索引アータとして上記案引作成部 1 9 にてもちれる。この案引作成部 1 9 はこれらのアータに基づいて、一つの記事内容を示す案引情報、例えばキーワード等を作成するものであり、これによって作成された上記案引情報は案引フィイル 2 1 に登録されるよりになっている。尚、この案引情報は、画像であってもよく、あるい

- 12 -

なせるので、その場合には出力文 排画像ファイル25 は表示出力となる。

第4図はこのように構成された本システムの 処理フローを示すものであり、 部 5 図(a)~(1)は その処理形態を模式的に示したものである。第 5 図(1)は入力文書画像(原画)の一例を示すも のである。しかしてとの入力文積画像は、先ず 第5図的に示すように、一定長さ以上の凝線 (区画線)検出がなされたのち、とれを同図(6) に示すように細線表示して凝区画線位置が認識 される。認識したとれらの細線は計算機内部で は端点の座標としての数値情報として保持され る。しかるのち、入力文替画像は第5図(d)に示 すように一定長さ以上の横線検出に供される。 この場合、たとえば見出し文字中の横線分も多 く検出されることがある。しかしながら、との ような見出し中の線分は区面線としては不要な ものであり、従って、一定の段間隔以外の横線 消去や区画線の太さ以上の横線消去等の処理に より、第5図(e)に示すように不要な機線分を除 去する。写真領域でもそのよりな機械分が出る ととがあり、また様区画線抽出においても同様 のととが起こるが、その除去は同様にできる。 しかるのち残された横線を細線表示して横方向面 線位置が認識される。とれらの細線も縦方向の 細線と同様、計算機内部では線の端点の座標に 報として保持される。第5回(1)は上記の処理に よって認識された縦横の区画線を重ねて表示した たもので、とれによって記事文章の区画が示される。

-- 15--

施例のシステム構成を示すものである。

新聞・雑誌等の面像は光ディスク等の大容置の文書画像ファイル装置30に密視される。との書報画像は、元の紙面そのものの面像でもいるでもよく、また先に説明したシステムにもいいて 編集されたものであってを形式の設当記事の相付領域を示す領域構造情報、更には見出し成成されるの構造情報は磁気ディスク等によって構成されるファイル装置31にそれぞれファイルされる。

しかして今、入力部32に検索質問文や会話的な指示情報が与えられると、このマッチング部33が応答する。このマッチングの第33は、上記検索質問文を前記索引ファイル設置31上の文字コードで示される検索索引とを適いますング処理したり、あるいは入力データを適像内領域切出しパターンとして、あるのでなり、更になり、して、ののではない、更になり、しているのではない、更になり、しているのでは、更になり、しているのではない、更になり、しているのではない。

それを利用して行名はよい。その後、とのようにして特出して行名はよい。その後、5 図(1)に行見出しに従って第5 図(1)に同日というに見出したがまれば、1)に示すように見出しなが5 図(1)に示すように別になかが5 図(1)に示すように別になかが5 図(1)に示すように別ななが5 の関係から要なにしたというに対してでする。との編集された見出しや符号による案引用の情報ファイルにするとともで行加して問報ファイルにするとともで表してである。

かくしてこのような処理を実行することによって、所望とするある一つの記事内容のみまたは 紙面中の 指定領域のみを新聞紙面中から的 確に 切出し、且つデータ圧縮してファイル化することが可能となる。しかもこのファイル化処理に際して即時的な人為的判断の必要がないので、システムを実現するのが容易である。

ところで、本システムでは情報検索して表示 出力することも容易である。第6図は、その実

-16-

は 画像の特徴パラメータを抽出してマッチング処理等を行う。 このマッチング処理の結果、検索質問に該当する記事の件数等の統計的情報が計算部34にて求められて前記入力部32に与えられる。 この応答を受けて、入力部32を介して会話的に検索の実行が進められる。

ファイル装置31中に紙面内の割付領域を示式、文書を指すっていた。 20には、全紙面をその面の情報ファイルなけっては、全紙面ののはないない。 20には、 2

ところで、画像情報の検索効率向上を図る為には、学習機能を導入することが望ましい。即ち、学習機能 3 8 を設け、検索質問、マッチンク処理の結果、および原画像との対応関係を分析して、学習を行わしめるようにすればよい。このようにすれば、学習結果によって検索索引ファイルに個人的を情報が追加されたりして、そのファイルが学習の進行に従って再構成され、

-19-

とのような情報検索システムは第3図のファイル作成システムと一体化したものであってよい。またそれぞれ専用化して分離したものであっても、本発明の効果を十分発揮できることは云うまでもない。

ところで、
画像の為の似子ファイルとして光
アイスクを用いれば、その大容量などを用
いて、例えば新聞に事を画像のまま蓄戦であると
して、例えば新聞になる。
でないながであると
がでで、ないながののはがのかない。
とこれがでないながでない。
はないながでない。
はないないがでは、ないのではない。
とこれができるが、ないのではない。
とこれができるが、ないのではない。
とこれができるが、はいいのではない。
とこれができるが、はいいのではない。

以上、本システムにつき説明したように、本システムによれば不定形式の文韻であって、且

またシステムにおける学習に際しては、シソーラス部39を散けて、その検索シソーラスを作成したり、索引情報の間での連想情報を梅樂するようにしてもよい。このようにシソーラスを組立てることによって関連記事の検索が容易となると云う効果が突せられる。

- 20 -

つ文字コード化がなされてなく、しかも図表や 写真等を含む新聞や雑誌等の文書であっても、 会話的な環境で略々自動的に所望とする記事を 切出してファイル化したり、また検索の対象と して出力することができる。従って従来のよう に、記事の切出しや索引情報の付加に人手を煩 らわすととがなく、大幅な省力化を図り得る。 勿論パターン理解的な処理は100パーセント 確実に正しい結果を与えるようにはできない。 しかしながら本システムでは、機械的な切出し が不確かな部分は、一応最も確からしい切出し を行っておき、その上にその不確かな部分につ いての警告を記述するフラグを立てると共に、 テーブルにまとめておけば、後でオペレータが 一括してその部分についての処理を再検討する というパターン理解におけるポストエディット の手法を用いるととができる。とのような不確 かな記事は定位置に本登録せずに仮のメモリに 入れ、人間に対して検酎を求めるための警告情 報の内容は、例えば新聞の場合には上部欄外の

特開昭57-199066 (ア)

ペーシ、日付等の情報と、そのペーシ内での該当個所の概略位置とでよい。 切出しが不確かな部分は少ないので、 このようにすれば、人手の介入を非常に少なくしておきながら、実用的なシステムが構築できる。

また対象となる文書としては、新闻、雑誌、 審箱のみならず、辞書や百科事典等を含ませめ ととができ、これらの情報の蓄積、検索を始め として古文書のデータベース等が可能となる。 また特に日本語文章の場合、タイプライタによ るコード化に非常に手間がかかることから、日 本語情報処理に本システムを適用した場合、多 大な貢献をなし得、実用的利点が絶大である。

尚、本発明は上記実施例に限定されるものではない。例えば記事内容の境界検出を行間,段間スペースから検出して行うようにしてもよく、また見出しや写真等の検出を文字枠の大きさ等によって行うようにしてもよい。要するに本発明はその要旨を逸脱しない範囲で種々変形して実施することができる。

- 23 -

編集システム、 3 7 … 出力文群画像、 3 8 … 学 習機能、 3 9 … シソーラス部

出願人代理人 弁理士 鈴 江 武 彦

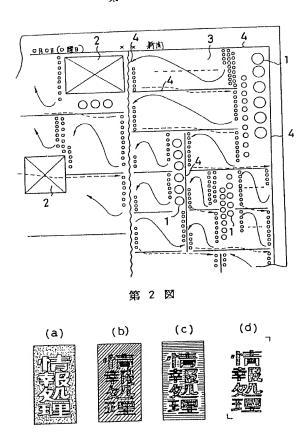
4. 図面の簡単な説明

第1 図は新聞記事の一例を示す図、第2図(a) ~(d) は見出しとその文字抽出形態を示す図、第3図は本発明に係る実施例システムの概略的な構成図、第4図は本システムにおける処理フローの一例を示す図、第5図(a) ~(1) は新聞記事の切抜きファイル化を示す図、第6図は本システムを用いた検索システムの概略的な構成図である。

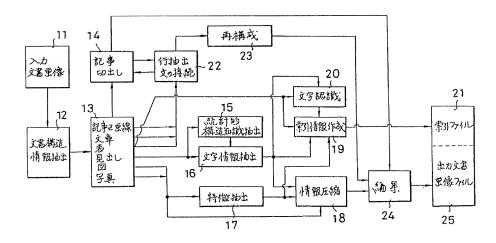
1 … 見出し、2 … 写真、3 … 配事本文、4 … 区画線、1 1 … 入力文神画像、1 2 … 文書構造情報抽出部、1 3 … 情報区分部、1 4 … 記事切出し部、1 5 … 構造知識抽出部、1 6 … 文字情報抽出部、1 7 … 特徵抽出部、1 8 … 情報圧縮部、1 9 … 索引作成部、2 0 … 文字認識部、2 1 … 索引ファイル、2 2 … 構造認識部、2 3 … 再構成部、2 4 … 編集部、2 5 … 出力文書画像ファイル技置、3 1 … ファイル装置、3 2 … 入力部、3 3 … マッチング部、3 4 … 計算部、3 5 … 画像検索部、3 6 …

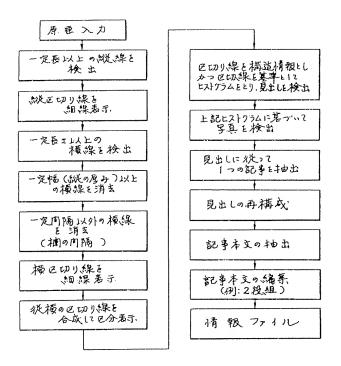
- 24 -

第1図

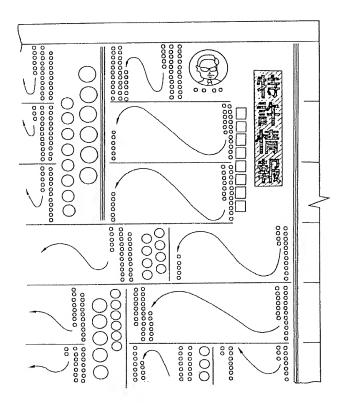


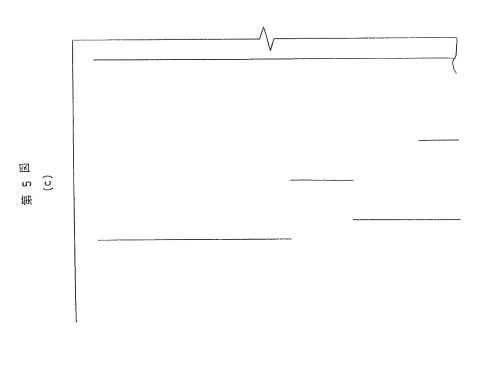
第 3 図

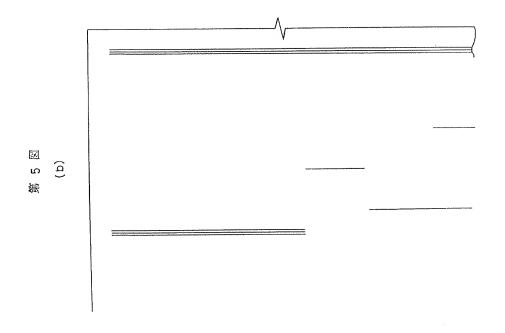


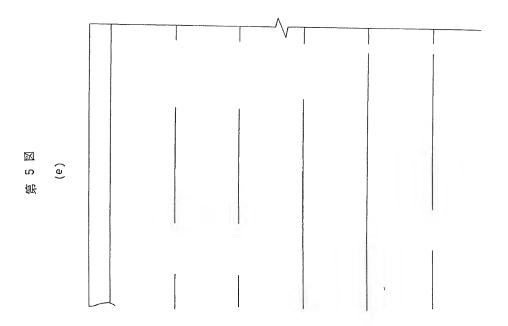


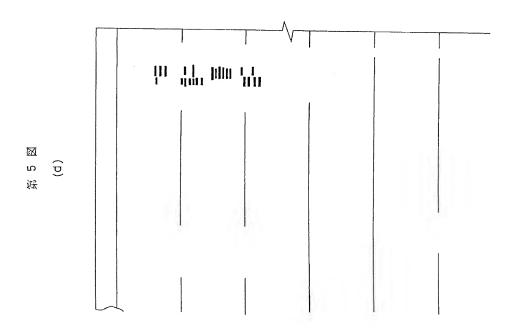
第 5 図 (a)

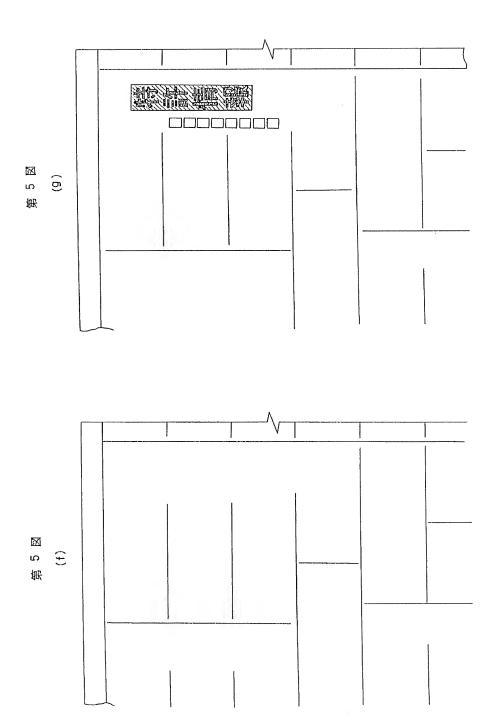


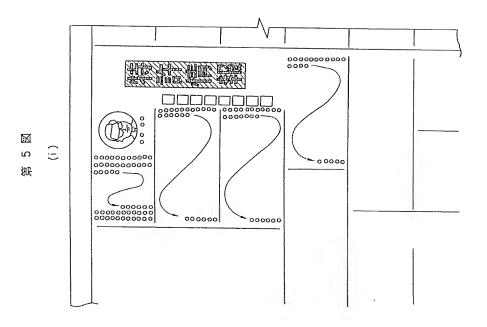


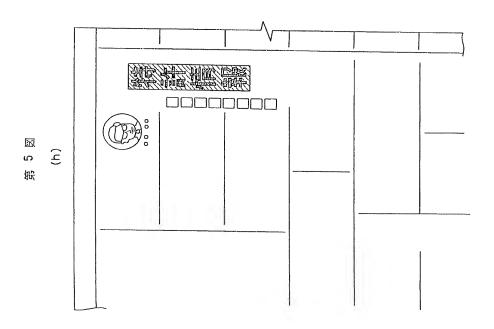


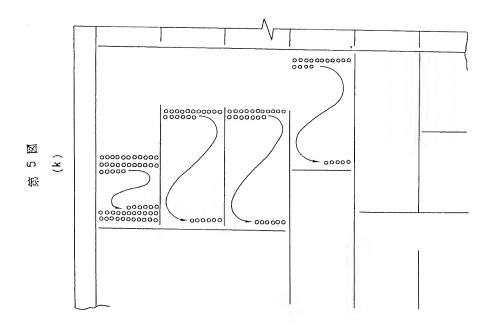


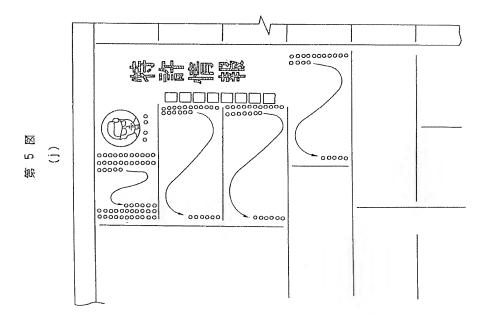






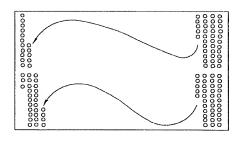




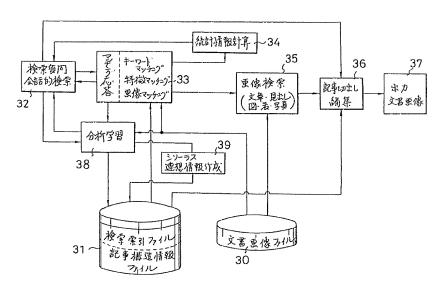


第 5 図

(l)



第 6 図



手 続 補 正 書

特許庁長官 島田春樹 殿

1. 事件の表示

特願昭 5 6 - 8 4 5 8 9 号

2. 発明の名称

新聞・雑誌等の切抜きフアイル作成システム

3. 補正をする者

事件との関係 特許出額人

坂 井 利 之

4. 代 理 人

理 人 住所 東京都港区成ノ門1丁目26番5号 第17章ビル 〒105 電話 03 (502) 3 1 8 1 (大代表) 震袋が 氏名 (5847) 弁理士 鈴 江 武 彦ピンジュ 「日本」

- 5. 自発補正
- 6. 補正の対象 明 細 書 お よ び 図 面

7. 補正の内容

- (1) 明細書、第 4 頁第 2 行目に「行間スペース」 とあるを「段間スペース」と訂正する。
- (2) 同、第7頁第11行目に「(Shrin king)」 とあるを「(Shrinking)」と訂正する。
- (3) 図面、第4図を別紙の通り訂正する。

2

才 4 図

